On Lipschitz Continuity and Smoothness of Loss Functions in Learning to Rank

Ambuj Tewari and Sougata Chaudhuri
{tewaria,sougata}@umich.edu

May 7, 2014

Abstract

In binary classification and regression problems, it is well understood that Lipschitz continuity and smoothness of the loss function play key roles in governing generalization error bounds for empirical risk minimization algorithms. In this paper, we show how these two properties affect generalization error bounds in the learning to rank problem. The learning to rank problem involves vector valued predictions and therefore the choice of the norm with respect to which Lipschitz continuity and smoothness are defined becomes crucial. Choosing the ℓ_∞ norm in our definition of Lipschitz continuity allows us to improve existing bounds. Furthermore, under smoothness assumptions, our choice enables us to prove rates that interpolate between $1/\sqrt{n}$ and $1/n$ rates. Application of our results to ListNet, a popular learning to rank method, gives state-of-the-art performance guarantees.

1 Introduction

In the setting of binary classification or regression, it is well known that Lipschitz continuity of the loss function impacts the generalization error of algorithms that minimize the loss on training examples. A key result that controls this impact is the Lipschitz contraction property of Rademacher (or Gaussian) complexity that, in turn, follows from the celebrated Ledoux-Talagrand contraction principle. It is also well known that Lipschitz continuity of the *derivative* of the loss, sometimes referred to as "smoothness", also impacts generalization error bounds. For instance, under smoothness, one can derive rates that interpolate between an "optimistic" $O(1/n)$ rate and a "pessimistic" $O(1/\sqrt{n})$ rate depending on whether or not the expected loss of the best predictor is close to zero.

In this paper, we investigate the impact of Lipschitz continuity and smoothness of loss function in the learning to rank problem. In learning to rank, the loss function takes a *vector* of predictions (or scores) as an argument. This leads to an interesting question that *does not* arise in binary classification or regression: which norm do we use to define Lipschitz continuity or smoothness of the loss function? Previous work has considered the use of the "default" Euclidean (or ℓ_2) norm for this purpose. We show that this choice can lead to suboptimal bounds and that better bounds can be obtained by using the ℓ_∞ norm in defining Lipschitz continuity and smoothness.

Using online regret bounds as a guide, we first show why one should expect to get better bounds under Lipschitz continuity with respect to the ℓ_∞ norm. However, online regret bounds require convexity of the loss function and, even under convexity, they do not establish uniform convergence of empirical loss averages to their expectations (and therefore do not lead to generalization error

bounds for empirical risk minimization (ERM)). Our first key result (Theorem 4) establishes a generalization error bound – via uniform convergence – for ERM under Lipschitz continuity of the loss. We consider linear scoring functions, a popular choice in theory as well as in practice. We consider both ℓ_2-norm and ℓ_1-norm bounded linear predictors. Our result in the latter case appears to be the first of its kind for learning to rank and can be useful if the dimensionality of the feature space is high and there is a need for feature selection.

Next we consider smoothness of the loss function, again with respect to the ℓ_∞ norm, and show why it is natural to expect that it is the right notion to derive rates that interpolate between optimistic and pessimistic cases. Our second key result (Theorem 9) is a generalization bound for ERM under smoothness. This is proved via a uniform convergence analysis using local Rademacher complexities. Not only was such a result not known for general, possibly non-convex, loss functions, we are not even aware of such a result for any specific loss function used in learning to rank.

As an illustration, we apply our key results to ListNet, a loss very popular in the learning to rank literature[1]. We discover that *both* its Lipschitz constant as well as smoothness constant *do not* increase with the number of documents being ranked per query. Our results, therefore, additionally provide novel theoretical insights into a popular learning to rank method.

2 Preliminaries

The increasing use of machine learning for web ranking and information retrieval tasks has led to a lot of recent research activity on the learning to rank problem (sometimes also called "subset ranking" to distinguish it from other related problems, for example, bipartite ranking). A training example in the learning to rank setting is of the form $((q, d_1, \ldots, d_m), y)$. Here q is a search query and $d_1, \ldots, d_m$ are m documents with varying degrees of *relevance* to the query. Human labelers provide the relevance vector $y \in \mathbb{R}^m$ where the entries in y contain the relevance labels for the m individual documents. Typically, y has integer-valued entries in the range $\{0, \ldots, Y_{\max}\}$ where $Y_{\max}$ is often less than 5. For our theoretical analysis, we get rid of some of these details by assuming that some feature map Ψ exists to map a query document pair (q, d) to $\mathbb{R}^d$. As a result, the training example $((q, d_1, \ldots, d_m), y)$ gets converted into (X, y) where $X = [\Psi(q, d_1), \ldots, \Psi(q, d_m)]^\top$ is an $m \times d$ matrix with the m query-document feature vector as rows. With this abstraction, we have an input space $\mathcal{X} \subseteq \mathbb{R}^{m \times d}$ and a label space $\mathcal{Y} \subseteq \mathbb{R}^m$.

A training set consists of iid examples $(X^{(1)}, y^{(1)}), \ldots, (X^{(n)}, y^{(n)})$ drawn from some underlying distribution D. To rank order the documents in a new instance $X \in \mathcal{X}$, often a score vector $s \in \mathbb{R}^m$ is computed. A ranking of the documents can then be obtained from s by sorting its entries in decreasing order, for instance. A common choice for the scoring function is to make it linear in the input X. Accordingly, we consider the following two classes in this paper:

$$\mathcal{F}_2 := \{X \mapsto Xw \ : \ X \in \mathbb{R}^{m \times d}, w \in \mathbb{R}^d, \|w\|_2 \leq W_2\},$$
$$\mathcal{F}_1 := \{X \mapsto Xw \ : \ X \in \mathbb{R}^{m \times d}, w \in \mathbb{R}^d, \|w\|_1 \leq W_1\}.$$

In the input space $\mathcal{X}$, it is natural to contain the rows of X to have a bound on the appropriate dual norm. Accordingly, whenever we use $\mathcal{F}_2$, the input space is set to

$$\mathcal{X} = \{X \in \mathbb{R}^{m \times d} \ : \ \forall j \in [m], \ \|X_j\|_2 \leq R_X\}$$

[1]The original ListNet paper has been cited close to 500 times already.

where X_j denotes jth row of X and $[m] := \{1, \ldots, m\}$. Similarly, when we use $\mathcal{F}_1$, we set

$$\mathcal{X} = \{X \in \mathbb{R}^{m \times d} \,:\, \forall j \in [m],\ \|X_j\|_\infty \le \bar{R}_X\}.$$

These are natural counterparts to the following function classes studied in binary classification and regression:

$$\mathcal{G}_2 := \{x \mapsto \langle x, w\rangle \,:\, x \in \mathbb{R}^d, \|x\|_2 \le R_X, w \in \mathbb{R}^d, \|w\|_2 \le W_2\},$$
$$\mathcal{G}_1 := \{x \mapsto \langle x, w\rangle \,:\, x \in \mathbb{R}^d, \|x\|_\infty \le \bar{R}_X, w \in \mathbb{R}^d, \|w\|_1 \le W_1\}.$$

A key ingredient in the basic setup of the learning to rank problem is a loss function $\phi : \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}_+$ where $\mathbb{R}_+$ denotes the set of non-negative real numbers. For vector valued scores, the Lipschitz constant of ϕ depends on the norm $|||\cdot|||$ that we decide to use in the score space:

$$\forall y \in \mathcal{Y}, s, s' \in \mathbb{R}^m,\ |\phi(s_1, y) - \phi(s_2, y)| \le G_\phi |||s_1 - s_2|||.$$

If ϕ is differentiable, this is equivalent to:

$$\forall y \in \mathcal{Y}, s \in \mathbb{R}^m,\ |||\nabla_s \phi(s, y)|||_\star \le G_\phi.$$

Similarly, the smoothness constant of ϕ depends on the norm used in the score space:

$$\forall y \in \mathcal{Y}, s, s' \in \mathbb{R}^m,\ |||\nabla_s \phi(s_1, y) - \nabla_s \phi(s_2, y)|||_\star \le H_\phi |||s_1 - s_2|||.$$

If ϕ is twice differentiable, this is equivalent to

$$\forall y \in \mathcal{Y}, s \in \mathbb{R}^m,\ |||\nabla_s^2 \phi(s, y)|||_{\text{op}} \le H_\phi$$

where $|||\cdot|||_{\text{op}}$ is the operator norm induced by the pair $|||\cdot|||, |||\cdot|||_\star$ and defined as $|||M|||_{\text{op}} := \sup_{v \ne 0} \frac{|||Mv|||_\star}{|||v|||}$. Define the expected loss of w under the distribution P as:

$$L_\phi(w) := \mathbb{E}_{(X,y)\sim D}\left[\phi(Xw, y)\right]$$

and its empirical loss on the sample as

$$\hat{L}_\phi(w) := \frac{1}{n}\sum_{i=1}^{n} \phi(X^{(i)} w, y^{(i)}).$$

We may occasionally refer to expectations w.r.t. the sample using $\widehat{\mathbb{E}}\,[\cdot]$. To reduce notational clutter, we often refer to (X, y) jointly by Z and $\mathcal{X} \times \mathcal{Y}$ by $\mathcal{Z}$.

2.1 Related work

Our work is directly motivated by a very interesting generalization bound for learning to rank due to Chapelle and Wu [2010, Theorem 1]. They considered a Lipschitz continuous loss ϕ with Lipschitz constant G_ϕ^{CW} w.r.t. *the ℓ_2 norm*. They show that, with probability at least $1 - \delta$,

$$\forall w \in \mathcal{F}_2,\ L_\phi(w) \le \hat{L}_\phi(w) + 3\, G_\phi^{CW} W_2 R_X \sqrt{\frac{m}{n}} + \sqrt{\frac{8 \log(1/\delta)}{n}}.$$

The dominant term on the right is $O(G_\phi^{CW} W_2 R_X \sqrt{m/n})$. Using the informal $\tilde{O}$ notation to hide logarithmic factors, our first key result (Theorem 4) will improve this to $\tilde{O}(G_\phi W_2 R_X/\sqrt{n})$ where G_ϕ is the Lipschitz constant of ϕ w.r.t. ℓ_∞ norm. Since $G_\phi \leq \sqrt{m} G_\phi^{CW}$, our bound can never be worse than their bound. However, as we show in Section 7, for the popular ListNet loss function, both G_ϕ and G_ϕ^{CW} are constants independent of m. In such cases, our bound offers an improvement by a factor of $\sqrt{m}$.

Our proof technique is very different from that of Chapelle and Wu [2010]. In the absence of an obvious contraction principle that would allow one to get rid of the loss function and work directly with the complexity of the underlying linear function class, they resorted to first principles and invoked Slepian's lemma. However, that forces them to define the Lipschitz constant w.r.t. the ℓ_2 norm. We deal with the absence of a general contraction principle by using covering number arguments that work quite nicely when the Lipschitz content is defined w.r.t. the ℓ_∞ norm.

To the best of our knowledge, our second key result (Theorem 9) has no direct predecessor in the learning to rank literature. But in terms of techniques, we do rely heavily on previous work by Bousquet [2002] and Srebro et al. [2010]. A key lemma (Lemma 6) we prove here is based on a vector extension of an inequality that was shown to hold in the scalar predictions case by Srebro et al. [2010] when a smooth loss function is used.

3 Online regret bounds under Lipschitz continuity

In this section, we build some intuition as to why we it is natural to use $\|\cdot\|_\infty$ in defining the Lipschitz constant of the loss ϕ. To this end, consider the following well known online gradient descent (OGD) regret guarantee. Recall that OGD refers to the simple online algorithm that makes the update $w_{t+1} \leftarrow w_t - \eta \nabla_w f_t(w_t)$ at time t. If we run OGD to generate w_t's, we have, for all $\|w\|_2 \leq W_2$:

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w) \leq \frac{W_2^2}{2\eta} + \eta G^2 T$$

where G is a bound on the maximum ℓ_2-norm of the gradients $\nabla_w f_t(w_t)$ and f_t's have to be *convex*. If $(X^{(1)}, y^{(1)}), \ldots, (X^{(n)}, y^{(n)})$ are iid then by setting $f_t(w) = \phi(X^{(t)} w, y^{(t)})$, $1 \leq t \leq n$, and using a standard online to batch conversion technique we can guarantee an excess risk bound of:

$$\forall \|w\|_2 \leq W_2,\ \mathbb{E}\left[L_\phi(\hat{w})\right] - L_\phi(w) \leq W_2 G \sqrt{\frac{2}{n}}$$

where $\hat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$ and G has to satisfy

$$G \geq \|\nabla_w f_t(w_t)\|_2 = \|(X^{(t)})^\top \nabla_s \phi(X^{(t)} w_t, y^{(t)})\|_2$$

where we use the chain rule to express ∇_w in terms of ∇_s. Finally, we can upper bound

$$\begin{aligned}\|(X^{(t)})^\top \nabla_s \phi(X^{(t)} w_t, y^{(t)})\|_2 &\leq \|(X^{(t)})^\top\|_{1\to 2} \cdot \|\nabla_s \phi(X^{(t)} w_t, y^{(t)})\|_1 \\ &= \max_{j=1}^{m} \|X_j\|_2 \cdot \|\nabla_s \phi(X^{(t)} w_t, y^{(t)})\|_1 \leq R_X \|\nabla_s \phi(X^{(t)} w_t, y^{(t)})\|_1 \quad (1)\end{aligned}$$

because of the following lemma.

Lemma 1. *For any* $1 \leq p \leq \infty$,

$$\|X^\top\|_{1\to p} = \|X\|_{q\to\infty} = \max_{j=1}^{m} \|X_j\|_p \ ,$$

where q *is the dual exponent of* p *(i.e.,* $\frac{1}{q} + \frac{1}{p} = 1$*).*

Proof. The first equality is true because

$$\begin{aligned}\|X^\top\|_{1\to p} &= \sup_{v\neq 0} \frac{\|X^\top v\|_p}{\|v\|_1} = \sup_{v\neq 0}\sup_{u\neq 0} \frac{\langle X^\top v, u\rangle}{\|v\|_1\|u\|_q} \\ &= \sup_{u\neq 0}\sup_{v\neq 0} \frac{\langle v, Xu\rangle}{\|v\|_1\|u\|_q} = \sup_{u\neq 0} \frac{\|Xu\|_\infty}{\|u\|_q} = \|X\|_{q\to\infty}.\end{aligned}$$

The second is true because

$$\begin{aligned}\|X\|_{q\to\infty} &= \sup_{u\neq 0} \frac{\|Xu\|_\infty}{\|u\|_q} = \sup_{u\neq 0} \max_{j=1}^{m} \frac{|\langle X_j, u\rangle|}{\|u\|_q} \\ &= \max_{j=1}^{m}\sup_{u\neq 0} \frac{|\langle X_j, u\rangle|}{\|u\|_q} = \max_{j=1}^{m} \|X_j\|_p.\end{aligned}$$

□

Thus, we have shown that if ϕ has Lipschitz constant G_ϕ w.r.t. $\|\cdot\|_\infty$, then we can guarantee an $O(G_\phi W_2 R_X/\sqrt{n})$ excess risk bound. This is encouraging but there are two deficiencies of this approach based on online regret bounds. First, there is no way to generalize the result to Lipschitz, but *non-convex* loss functions. Second, the result applies to the output of a specific algorithm. That is, we do not get uniform convergence bounds or excess risk bounds for ERM. We now address these issues.

4 Generalization error bounds under Lipschitz continuity

The above discussion suggests that we have a possibility of deriving tighter, possibly m-independent, generalization error bounds by assuming that ϕ is Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. The standard approach in binary classification is to appeal to the Ledoux-Talagrand contraction principle for Rademacher complexity [Bartlett and Mendelson, 2003] by getting rid of the Lipschitz loss function (that takes a scalar argument in the binary classification case) and incurring a factor equal to the Lipschitz constant of the loss in the Rademacher complexity bound. It is not immediately clear how such an approach would work when the loss takes vector valued arguments and is Lipschitz w.r.t. $\|\cdot\|_\infty$ since we are not aware of an appropriate extension of the Ledoux-Talagrand contraction principle. Note that Lipschitz continuity w.r.t. the Euclidean norm $\|\cdot\|_2$ does not pose a significant challenge since Slepian's lemma can be applied to get rid of the loss function. As we mentioned before, several authors have already exploited Slepian's lemma in this context [Bartlett and Mendelson, 2003, Chapelle and Wu, 2010].

In the absence of a general principle that would allow us to deal with an arbitrary loss function that is Lipschitz w.r.t. $\|\cdot\|_\infty$, we take a route involving covering numbers. Define the data-dependent (pseudo-)metric:

$$d_\infty^{Z^{(1:n)}}(w, w') := \max_{i=1}^{n} \left|\phi(X^{(i)}w, y^{(i)}) - \phi(X^{(i)}w', y^{(i)})\right|$$

and let $\mathcal{N}_\infty(\epsilon, \mathcal{F}, Z^{(1:n)})$ be the covering number at scale ϵ of the class $\mathcal{F} = \mathcal{F}_1$ or $\mathcal{F}_2$ w.r.t. the above metric. Also define

$$\mathcal{N}_\infty(\epsilon, \mathcal{F}, n) := \max_{Z^{(1:n)}} \mathcal{N}_\infty(\epsilon, \mathcal{F}, Z^{(1:n)}).$$

With these definitions in place, we can state our first result on covering numbers.

Proposition 2. *Let the loss ϕ be Lipschitz in its first argument w.r.t. $\|\cdot\|_\infty$ with constant G_ϕ. Then the following covering number bounds hold:*

$$\log_2 \mathcal{N}_\infty(\epsilon, \mathcal{F}_2, n) \leq \left\lceil \frac{G_\phi^2 W_2^2 R_X^2}{\epsilon^2} \right\rceil \log_2(2mn+1),$$

$$\log_2 \mathcal{N}_\infty(\epsilon, \mathcal{F}_1, n) \leq \left\lceil \frac{288\, G_\phi^2 W_1^2 \bar{R}_X^2 (2+\log d)}{\epsilon^2} \right\rceil \log_2 \left(2 \left\lceil \frac{8 G_\phi W_1 \bar{R}_X}{\epsilon} \right\rceil mn + 1 \right).$$

Proof. Note that

$$\max_{i=1}^n \left| \phi(X^{(i)} w, y^{(i)}) - \phi(X^{(i)} w', y^{(i)}) \right| \leq G_\phi \cdot \max_{i=1}^n \max_{j=1}^m \left| \left\langle X_j^{(i)}, w \right\rangle - \left\langle X_j^{(i)}, w' \right\rangle \right|.$$

This immediately implies that if we have a cover of the class $\mathcal{G}_2$ (respectively $\mathcal{G}_1$) at scale ϵ/G_ϕ w.r.t. the metric

$$\max_{i=1}^n \max_{j=1}^m \left| \left\langle X_j^{(i)}, w \right\rangle - \left\langle X_j^{(i)}, w' \right\rangle \right|$$

then it is also a cover of $\mathcal{F}_2$ (respectively $\mathcal{F}_1$) w.r.t. $d_\infty^{Z^{(1:n)}}$. From the point of view of the scalar valued function classes $\mathcal{G}_1$ or $\mathcal{G}_2$, the vectors $X_j^{(i)}$ constitute a data set of size mn. Therefore, we have

$$\mathcal{N}_\infty(\epsilon, \mathcal{F}_2, n) \leq \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_2, mn) \tag{2}$$

as well as

$$\mathcal{N}_\infty(\epsilon, \mathcal{F}_1, n) \leq \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_1, mn). \tag{3}$$

Now we appeal to the following bound due to Zhang [2002, Corollary 3 and Corollary 5]:

$$\log_2 \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_2, mn) \leq \left\lceil \frac{G_\phi^2 W_2^2 R_X^2}{\epsilon^2} \right\rceil \log_2(2mn+1)$$

$$\log_2 \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_1, mn) \leq \left\lceil \frac{288\, G_\phi^2 W_1^2 \bar{R}_X^2 (2+\ln d)}{\epsilon^2} \right\rceil \log_2 \left(2 \lceil 8 G_\phi W_1 \bar{R}_X / \epsilon \rceil mn + 1 \right)$$

Plugging these into (2) and (3) respectively proves the result. □

Recall that the covering number $\mathcal{N}_2(\epsilon, \mathcal{F}, Z^{(1:n)})$ uses the (pseudo-)metric:

$$d_2^{Z^{(1:n)}}(w, w') := \left(\sum_{i=1}^n \frac{1}{n} \left(\phi(X^{(i)} w, y^{(i)}) - \phi(X^{(i)} w', y^{(i)}) \right)^2 \right)^{1/2}$$

It is well known that a control on $\mathcal{N}_2(\epsilon, \mathcal{F}, Z^{(1:n)})$ provides control on the empirical Rademacher complexity and that $\mathcal{N}_2$ covering numbers are smaller than $\mathcal{N}_\infty$ ones. For us, it will be convenient

to use a more refined version[2] due to Mendelson [2002]. Let $\mathcal{F}$ be a class of functions uniformly bounded by B. Then, we have

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \leq \inf_{\alpha>0}\left(4\alpha + 10\int_{\alpha}^{\sup_{f\in\mathcal{F}}\sqrt{\widehat{\mathbb{E}}[f^2]}}\sqrt{\frac{\log_2 \mathcal{N}_2(\epsilon,\mathcal{F},Z^{(1:n)})}{n}}d\epsilon\right) \tag{4}$$

$$\leq \inf_{\alpha>0}\left(4\alpha + 10\int_{\alpha}^{B}\sqrt{\frac{\log_2 \mathcal{N}_2(\epsilon,\mathcal{F},Z^{(1:n)})}{n}}d\epsilon\right). \tag{5}$$

Here $\widehat{\mathfrak{R}}_n(\mathcal{F})$ is the empirical Rademacher complexity of the class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ defined as

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) := \mathbb{E}_{\sigma_{1:n}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right],$$

where $\sigma_{1:n} = (\sigma_1,\ldots,\sigma_n)$ are iid Rademacher (symmetric Bernoulli) random variables.

Corollary 3. *Let ϕ be Lipschitz w.r.t. $\|\cdot\|_\infty$ and uniformly bounded[3] by B for $w \in \mathcal{F}_2$ (or $\mathcal{F}_1$ as the case may be). Then the empirical Rademacher complexities of the classes $\mathcal{F}_2, \mathcal{F}_1$ are bounded as*

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_2) \leq 10 G_\phi W_2 R_X \sqrt{\frac{\log_2(3mn)}{n}}\log\frac{6B\sqrt{n}}{5G_\phi W_2 R_X\sqrt{\log_2(3mn)}},$$

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_1) \leq 120\sqrt{2}G_\phi W_1 \bar{R}_X\sqrt{\frac{\log(d)\ \log_2(24mnG_\phi W_1\bar{R}_X)}{n}} \times \log^2\frac{B+24mnG_\phi W_1\bar{R}_X}{40\sqrt{2}G_\phi W_1\bar{R}_X\sqrt{\log(d)\ \log_2(24mnG_\phi W_1\bar{R}_X)}}.$$

Proof. These follow by simply plugging in estimates from Proposition 2 into (5) and choosing α optimally. □

Control on the Rademacher complexity immediately leads to uniform convergence bounds and generalization error bounds for ERM. The informal $\tilde{O}$ notation hides factors logarithmic in $m, n, B, G_\phi, R_X, W_1, W_2$. Note that all hidden factors are small and computable from the results above.

Theorem 4. *Suppose ϕ is Lipschitz w.r.t. $\|\cdot\|_\infty$ with constant G_ϕ and is uniformly bounded by B over the function class being used. With probability at least $1-\delta$,*

$$\forall w \in \mathcal{F}_2,\ L_\phi(w) \leq \hat{L}_\phi(w) + \tilde{O}\left(G_\phi W_2 R_X\sqrt{\frac{1}{n}} + B\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

and therefore with probability at least $1-2\delta$,

$$L_\phi(\hat{w}) \leq L_\phi(w^\star) + \tilde{O}\left(G_\phi W_2 R_X\sqrt{\frac{1}{n}} + B\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

[2] We use a further refinement due to Srebro and Sridharan available at `http://ttic.uchicago.edu/~karthik/dudley.pdf`

[3] A uniform bound on the loss easily follows under the (very reasonable) assumption that $\forall y, \exists s_y$ s.t. $\phi(s_y, y) = 0$. Then $\phi(Xw, y) \leq G_\phi\|Xw - s_y\|_\infty \leq G_\phi(W_2 R_X + \max_{y\in\mathcal{Y}}\|s_y\|_\infty)$.

where $\hat{w}$ is an empirical risk minimizer (i.e. a minimizer of $\hat{L}_\phi(w)$) over $\mathcal{F}_2$. The same result holds for the class $\mathcal{F}_1$ with $G_\phi W_2 R_X$ replaced with $G_\phi W_1 \bar{R}_X \sqrt{\log(d)}$.

Proof. Follows from standard bounds using Rademacher complexity. See, for example, Bartlett and Mendelson [2003]. □

As we said before, ignoring logarithmic factors, the bound for $\mathcal{F}_2$ is an improvement over the bound of Chapelle and Wu [2010]. The generalization bound for $\mathcal{F}_1$ appears to be new and could be useful in learning to rank situations involving high dimensional features.

5 Online regret bounds under smoothness

Let us go back to OGD guarantee, this time presented in a slightly more refined version. If we run OGD with learning rate η then, for all $\|w\|_2 \leq W_2$:

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w) \leq \frac{W_2^2}{2\eta} + \eta \sum_{t=1}^{T} \|g_t\|_2^2$$

where $g_t = \nabla_w f_t(w_t)$ (if f_t is not differentiable at w_t then we can set g_t to be an arbitrary subgradient of f_t at w_t). Now assume that all f_t's are non-negative functions and are smooth w.r.t. $\|\cdot\|_2$ with constant H. Lemma 3.1 of Srebro et al. [2010] tells us that any non-negative, smooth function $f(w)$ enjoy an important *self-bounding* property for the gradient:

$$\|\nabla_w f_t(w)\|_2 \leq \sqrt{4H f_t(w)}$$

which bounds the magnitude of the gradient of f at a point in terms of the value of the function itself at that point. This means that $\|g_t\|_2^2 \leq 4H f_t(w_t)$ which, when plugged into the OGD guarantee, gives:

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w) \leq \frac{W_2^2}{2\eta} + 4\eta H \sum_{t=1}^{T} f_t(w_t)$$

Again, setting $f_t(w) = \phi(X^{(t)} w, y^{(t)})$, $1 \leq t \leq n$, and using the online to batch conversion technique, we can arrive at the bound: for all $\|w\|_2 \leq W_2$:

$$\mathbb{E}\left[L_\phi(\hat{w})\right] \leq \frac{L_\phi(w)}{(1 - 4\eta H)} + \frac{W_2^2}{2\eta(1 - 4\eta H)n}$$

At this stage, we can fix $w = w^\star$, the optimal ℓ_2-norm bounded predictor and optimize the right hand side over η by setting

$$\eta = \frac{W_2}{4HW_2 + 2\sqrt{4H^2W_2^2 + 2HL_\phi(w^\star)n}}. \tag{6}$$

After plugging this value of η in the bound above and some algebra (see Section A), we get the upper bound

$$\mathbb{E}\left[L_\phi(\hat{w})\right] \leq L_\phi(w^\star) + \sqrt{\frac{2HW_2^2 L_\phi(w^\star)}{n}} + \frac{8HW_2^2}{n}. \tag{7}$$

Such a rate interpolates between a $1/\sqrt{n}$ rate in the "pessimistic" case ($L_\phi(w^\star) > 0$) and the $1/n$ rate in the "optimistic" case ($L_\phi(w^\star) = 0$) (this terminology is due to Panchenko [2002]).

We have not yet related the smoothness constant H to the smoothness of the underlying loss ϕ (views as a function of the score vector). We do this now. Assuming ϕ to be twice differentiable. Then we need to choose H such that

$$H \geq \|\nabla_w^2 \phi(X^{(t)}w, y^{(t)})\|_{2\to 2} = \|X^\top \nabla_s^2 \phi(X^{(t)}w, y^{(t)})X\|_{2\to 2}$$

using the chain rule to express ∇_w^2 in terms of ∇_s^2. Note that, for OGD, we need smoothness in w w.r.t. $\|\cdot\|_2$ which is why the matrix norm above is the operator norm corresponding to the pair $\|\cdot\|_2, \|\cdot\|_2$. In fact, when we say "operator norm" without mentioning the pair of norms involved, it is this norm that is usually meant. It is well known that this norm is equal to the largest singular value of the matrix. But, just as before, we can bound this in terms of the smoothness constant of ϕ w.r.t. $\|\cdot\|_\infty$:

$$\begin{aligned}\|(X^{(t)})^\top \nabla_s^2 \phi(X^{(t)}w, y^{(t)})X^{(t)}\|_{2\to 2} &\leq \|(X^{(t)})^\top\|_{1\to 2} \cdot \|\nabla_s^2 \phi(X^{(t)}w, y^{(t)})\|_{\infty\to 1} \cdot \|X^{(t)}\|_{2\to\infty} \\ &\leq \left(\max_{j=1}^m \|X_j^{(t)}\|\right)^2 \cdot \|\nabla_s^2 \phi(X^{(t)}w, y^{(t)})\|_{\infty\to 1} \leq R_X^2 \|\nabla_s^2 \phi(X^{(t)}w, y^{(t)})\|_{\infty\to 1}. \quad (8)\end{aligned}$$

where we used Lemma 1 once again.

This result using online regret bounds is great for building intuition but suffers from the two defects we mentioned at the end of Section 3. In the smoothness case, it additionally suffers from a more serious defect: the correct choice of the learning rate η requires knowledge of $L_\phi(w^\star)$ which is seldom available.

6 Generalization error bounds under smoothness

Once again, to prove a general result for possibly non-convex smooth losses, we will adopt an approach based on covering numbers. To begin, we will need the following useful lemma from Srebro et al. [2010, Lemma A.1 in the Supplementary Material]. Note that, for functions over the reals, we do not need to talk about the norm when dealing with smoothness since essentially the only norm available is the absolute value.

Lemma 5. *For any h-smooth non-negative function $f : \mathbb{R} \to \mathbb{R}_+$ and any $t, r \in \mathbb{R}$ we have*

$$(f(t) - f(r))^2 \leq 6h(f(t) + f(r))(t - r)^2.$$

We first provide an easy extension of this lemma to the vector case.

Lemma 6. *If $\phi : \mathbb{R}^m \to \mathbb{R}_+$ is a non-negative function with smoothness constant H_ϕ w.r.t. a norm $|||\cdot|||$ then for any $s_1, s_2 \in \mathbb{R}^m$ we have*

$$(\phi(s_1) - \phi(s_2))^2 \leq 6H_\phi \cdot (\phi(s_1) + \phi(s_2)) \cdot |||s_1 - s_2|||^2.$$

Proof. See Appendix B. □

Using the basic idea behind local Rademacher complexity analysis, we define the following loss class:

$$\mathcal{F}_{\phi,2}(r) := \{w \in \mathcal{F}_2 \ : \ \hat{L}_\phi(w) \leq r\}.$$

Note that this is a random subclass of functions since $\hat{L}_\phi(w)$ is a random variable.

Proposition 7. *Let ϕ be smooth, in its first argument, w.r.t. $\|\cdot\|_\infty$ with constant H_ϕ. The covering numbers of $\mathcal{F}_{\phi,2}(r)$ in the $d_2^{Z^{(1:n)}}$ metric defined above are bounded as follows:*

$$\log_2 \mathcal{N}_2(\epsilon, \mathcal{F}_{\phi,2}(r), Z^{(1:n)}) \leq \left\lceil \frac{12 H_\phi \, W_2^2 \, R_X^2 \, r}{\epsilon^2} \right\rceil \log_2(2mn+1).$$

Proof. Let $w, w' \in \mathcal{F}_{\phi,2}(r)$. Using Lemma 6

$$\begin{aligned}
&\sum_{i=1}^n \frac{1}{n} \left(\phi(X^{(i)} w, y^{(i)}) - \phi(X^{(i)} w', y^{(i)}) \right)^2 \\
&\leq 6 H_\phi \sum_{i=1}^n \frac{1}{n} \left(\phi(X^{(i)} w, y^{(i)}) + \phi(X^{(i)} w', y^{(i)}) \right) \|X^{(i)} w - X^{(i)} w'\|_\infty^2 \\
&\leq 6 H_\phi \cdot \max_{i=1}^n \|X^{(i)} w - X^{(i)} w'\|_\infty^2 \cdot \sum_{i=1}^n \frac{1}{n} \left(\phi(X^{(i)} w, y^{(i)}) + \phi(X^{(i)} w', y^{(i)}) \right) \\
&= 6 H_\phi \cdot \max_{i=1}^n \|X^{(i)} w - X^{(i)} w'\|_\infty^2 \cdot \left(\hat{L}_\phi(w) + \hat{L}_\phi(w') \right) \\
&\leq 12 H_\phi r \cdot \max_{i=1}^n \|X^{(i)} w - X^{(i)} w'\|_\infty^2.
\end{aligned}$$

where the last inequality follows because $\hat{L}_\phi(w) + \hat{L}_\phi(w') \leq 2r$.

This immediately implies that if we have a cover of the class $\mathcal{G}_2$ at scale $\epsilon / \sqrt{12 H_\phi r}$ w.r.t. the metric

$$\max_{i=1}^n \max_{j=1}^m \left| \left\langle X_j^{(i)}, w \right\rangle - \left\langle X_j^{(i)}, w' \right\rangle \right|$$

then it is also a cover of $\mathcal{F}_{\phi,2}(r)$ w.r.t. $d_2^{Z^{(1:n)}}$. Therefore, we have

$$\mathcal{N}_2(\epsilon, \mathcal{F}_{\phi,2}(r), Z^{(1:n)}) \leq \mathcal{N}_\infty(\epsilon / \sqrt{12 H_\phi r}, \mathcal{G}_2, mn). \tag{9}$$

Appealing once again to a result by Zhang [2002, Corollary 3], we get

$$\log_2 \mathcal{N}_\infty(\epsilon / \sqrt{12 H_\phi r}, \mathcal{G}_2, mn) \leq \left\lceil \frac{12 H_\phi \, W_2^2 \, R_X^2 \, r}{\epsilon^2} \right\rceil \log_2(2mn+1)$$

which finishes the proof. □

Corollary 8. *Let ϕ be smooth w.r.t. $\|\cdot\|_\infty$ and uniformly bounded by B for $w \in \mathcal{F}_2$. Then the empirical Rademacher complexity of the class $\mathcal{F}_{\phi,2}(r)$ is bounded as*

$$\widehat{\mathfrak{R}}_n \left(\mathcal{F}_{\phi,2}(r) \right) \leq 4 \sqrt{r} C \log \frac{3\sqrt{B}}{C}$$

where $C = 5\sqrt{3} W_2 R_X \sqrt{\frac{H_\phi \log_2(3mn)}{n}}$.

Proof. See Appendix C. □

With the above corollary in place we can now prove our second key result.

Theorem 9. *Suppose ϕ is smooth w.r.t. $\|\cdot\|_\infty$ with constant H_ϕ and is uniformly bounded by B over $\mathcal{F}_2$. With probability at least $1-\delta$,*

$$\forall w \in \mathcal{F}_2,\ L_\phi(w) \le \hat{L}_\phi(w) + \tilde{O}\left(\sqrt{\frac{L_\phi(w)D_0}{n}} + \frac{D_0}{n}\right)$$

where $D_0 = B\log(1/\delta) + W_2^2 R_X^2 H_\phi$. Moreover, with probability at least $1-2\delta$,

$$L_\phi(\hat{w}) \le L_\phi(w^\star) + \tilde{O}\left(\sqrt{\frac{L_\phi(w^\star)D_0}{n}} + \frac{D_0}{n}\right)$$

where $\hat{w}, w^\star$ are minimizers of $\hat{L}_\phi(w)$ and $L_\phi(w)$ respectively (over $w \in \mathcal{F}_2$).

Proof. We appeal to Theorem 6.1 of Bousquet [2002] that assumes there exists an upper bound

$$\widehat{\mathfrak{R}}_n\left(\mathcal{F}_{2,\phi}(r)\right) \le \psi_n(r)$$

where $\psi_n : [0,\infty) \to \mathbb{R}_+$ is a non-negative, non-decreasing, non-zero function such that $\psi_n(r)/\sqrt{r}$ is non-increasing. The upper bound in Corollary 8 above satisfies these conditions and therefore we set $\psi_n(r) = 4\sqrt{r}C\log\frac{3\sqrt{B}}{C}$ with C as defined in Corollary 8. From Bousquet's result, we know that, with probability at least $1-\delta$,

$$\forall w \in \mathcal{F}_2,\ L_\phi(w) \le \hat{L}_\phi(w) + 45r_n^\star + \sqrt{8r_n^\star L_\phi(w)} + \sqrt{4r_0 L_\phi(w)} + 20r_0$$

where $r_0 = B(\log(1/\delta) + \log\log n)/n$ and $r_n^\star$ is the largest solution to the equation $r = \psi_n(r)$. In our case, $r_n^\star = \left(4C\log\frac{3\sqrt{B}}{C}\right)^2$. This proves the first inequality

Now, using the above inequality with $w = \hat{w}$, the empirical risk minimizer and noting that $\hat{L}_\phi(\hat{w}) \le \hat{L}_\phi(w^\star)$, we get

$$L_\phi(\hat{w}) \le \hat{L}_\phi(w^\star) + 45r_n^\star + \sqrt{8r_n^\star L_\phi(\hat{w})} + \sqrt{4r_0 L_\phi(\hat{w})} + 20r_0$$

The second inequality now follows after some elementary calculations detailed in Appendix D. □

7 Application to ListNet

We now apply the results of this paper to the ListNet loss function [Lan et al., 2009]. ListNet is a popular learning method with competitive performance on a variety of benchmark data sets. It is defined in the following way[4]. Define m maps from $\mathbb{R}^m$ to $\mathbb{R}$ as: $P_j(v) = \exp(v_j)/\sum_{j=1}^m \exp(v_j)$ for $j \in [m]$. Then, we have

$$\phi_{\mathrm{LN}}(s,y) = -\sum_{j=1}^m P_j(y)\log P_j(s).$$

Since our results need the constants G_ϕ and H_ϕ we first compute them for the ListNet loss function.

[4]The ListNet paper actually defines a family of losses based on probability models for top k documents. We use $k=1$ in our definition since that is the version implemented in their experimental results.

Proposition 10. *The Lipschitz and smoothness constants of* ϕ_{LN} *w.r.t.* $\|\cdot\|_\infty$ *satisfy* $G_{\phi_{\text{LN}}} \leq 2$ *and* $H_{\phi_{\text{LN}}} \leq 2$ *for any* $m \geq 1$.

Proof. See Appendix E. □

Since the bounds above are independent of m, so the generalization bounds resulting from their use in Theorem 4 and Theorem 9 will also be independent of m (up to logarithmic factors). We are not aware of prior generalization bounds for ListNet that do not scale with the number of documents. In particular, the results of Lan et al. [2009] have an $m!$ dependence since they consider the top-m version of ListNet. However, even if the top-1 variant above is considered, it seems that their proof technique will result in at least a linear dependence on m and can never result in as tight a bound as we get from our general results. Moreover, generalization error bounds for ListNet that interpolate between the pessimistic $1/\sqrt{n}$ and optimistic $1/n$ rates have not been provided before.

8 Conclusion

In this paper, we derived generalization error bounds for learning to rank under Lipschitz continuity and smoothness assumptions on the loss function. Under the latter assumption, our bounds interpolate between $1/\sqrt{n}$ and $1/n$ rates. We showed why it is natural to measure Lipschitz and smoothness constants for learning to rank losses with respect to the ℓ_∞ norm. Our bounds under Lipschitz continuity improve previous results whereas our results under smoothness assumptions, to the best of our knowledge, are the first of their kind in the learning to rank setting.

A number of interesting avenues present themselves for further exploration. If the covering number approach can be by-passed via an argument directly at the level of Rademacher complexity, then it might be possible to avoid some logarithmic factors that we incur in our bounds. Another thing to note is that our arguments do not rely much on the specifics of the learning to rank setting and might apply more generally to situations, such as multi-label learning, that involve losses taking a vector of predictions as an argument.

Acknowledgments

We gratefully acknowledge the support of NSF under grant IIS-1319810.

References

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, PhD thesis, Ecole Polytechnique, 2002.

O. Chapelle and M. Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010.

Yanyan Lan, Tie-Yan Liu, Zhiming Ma, and Hang Li. Generalization analysis of listwise learning-to-rank algorithms. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 577–584, 2009.

Shahar Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *Information Theory, IEEE Transactions on*, 48(1):251–263, 2002.

Dmitriy Panchenko. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise, and fast rates. In *Advances in Neural Information Processing Systems 23*, pages 2199–2207, 2010.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.

A Calculations involved in deriving Equation (7)

Plugging in the value of η from (6) into the expression

$$\frac{L_\phi(w^\star)}{(1-4\eta H)} + \frac{W_2^2}{2\eta(1-4\eta H)n}$$

yields (using the shorthand $L^\star$ for $L_\phi(w^\star)$)

$$L^\star + \frac{2HW_2L^\star}{\sqrt{4H^2W_2^2+2HL^\star n}} + \frac{W_2}{n}\left[\frac{4H^2W_2^2}{\sqrt{4H^2W_2^2+2HL^\star n}} + \sqrt{4H^2W_2^2+2HL^\star n} + 4HW_2\right]$$

Denoting HW_2^2/n by x, this simplifies to

$$L^\star + \frac{2\sqrt{x}L^\star + 4x\sqrt{x}}{\sqrt{4x+2L^\star}} + \sqrt{x}\sqrt{4x+2L^\star} + 4x.$$

Using the arithmetic mean-geometric mean inequality to upper bound the middle two terms gives

$$L^\star + \sqrt{2xL^\star + 4x^2} + 4x.$$

Finally, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we get our final upper bound

$$L^\star + \sqrt{2xL^\star} + 8x.$$

B Proof of Lemma 6

Proof. Consider the function

$$f(t) = \phi((1-t)s_1 + ts_2).$$

It is clearly non-negative. Moreover

$$\begin{aligned}|f'(t_1) - f'(t_2)| &= |\left\langle \nabla_s \phi(s_1 + t_1(s_2 - s_1)) - \nabla_s \phi(s_1 + t_2(s_2 - s_1)), s_2 - s_1 \right\rangle| \\ &\leq |||\nabla_s \phi(s_1 + t_1(s_2 - s_1)) - \nabla_s \phi(s_1 + t_2(s_2 - s_1))|||_\star \cdot |||s_2 - s_1||| \\ &\leq H_\phi |t_1 - t_2| \, |||s_2 - s_1|||^2\end{aligned}$$

and therefore it is smooth with constant $h = H_\phi |||s_2 - s_1|||^2$. Appealing to Lemma 5 now gives

$$(f(1) - f(0))^2 \leq 6H_\phi |||s_2 - s_1|||^2 (f(1) + f(0))(1 - 0)^2$$

which proves the lemma since $f(0) = \phi(s_1)$ and $f(1) = \phi(s_2)$. □

C Proof of Corollary 8

Proof. We plug in Proposition 7's estimate into (4):

$$\begin{aligned}\widehat{\mathfrak{R}}_n\left(\mathcal{F}_{\phi,2}(r)\right) &\leq \inf_{\alpha > 0} \left(4\alpha + 10 \int_\alpha^{\sqrt{Br}} \sqrt{\frac{\left\lceil \frac{12 H_\phi W_2^2 R_X^2 \, r}{\epsilon^2} \right\rceil \log_2(2mn+1)}{n}} d\epsilon \right) \\ &\leq \inf_{\alpha > 0} \left(4\alpha + 20\sqrt{3} W_2 R_X \sqrt{\frac{r H_\phi \log_2(3mn)}{n}} \int_\alpha^{\sqrt{Br}} \frac{1}{\epsilon} d\epsilon \right) .\end{aligned}$$

Now choosing $\alpha = C\sqrt{r}$ where $C = 5\sqrt{3} W_2 R_X \sqrt{\frac{H_\phi \log_2(3mn)}{n}}$ gives us the upper bound

$$\widehat{\mathfrak{R}}_n\left(\mathcal{F}_{\phi,2}(r)\right) \leq 4\sqrt{r} C \left(1 + \log \frac{\sqrt{B}}{C} \right) \leq 4\sqrt{r} C \log \frac{3\sqrt{B}}{C}.$$

□

D Details of some calculations in the proof of Theorem 9

Using Bernstein's inequality, we have, with probability at least $1 - \delta$,

$$\begin{aligned}\hat{L}_\phi(w^\star) &\leq L_\phi(w^\star) + \sqrt{\frac{4 \mathrm{Var}[\phi(X w^\star, y)] \log(1/\delta)}{n}} + \frac{4B \log(1/\delta)}{n} \\ &\leq L_\phi(w^\star) + \sqrt{\frac{4B L_\phi(w^\star) \log(1/\delta)}{n}} + \frac{4B \log(1/\delta)}{n} \\ &\leq L_\phi(w^\star) + \sqrt{4 r_0 L_\phi(w^\star)} + 4r_0.\end{aligned}$$

Set $D_0 = 45 r_n^\star + 20 r_0$. Putting the two bounds together and using some simple upper bounds, we have, with probability at least $1 - 2\delta$,

$$\begin{aligned}L_\phi(\hat{w}) &\leq \sqrt{D_0 \hat{L}_\phi(w^\star)} + D_0, \\ \hat{L}_\phi(w^\star) &\leq \sqrt{D_0 L_\phi(w^\star)} + D_0.\end{aligned}$$

which implies that

$$L_\phi(\hat{w}) \le \sqrt{D_0}\sqrt{\sqrt{D_0 L_\phi(w^\star)} + D_0} + D_0.$$

Using $\sqrt{ab} \le (a+b)/2$ to simplify the first term on the right gives us

$$L_\phi(\hat{w}) \le \frac{D_0}{2} + \frac{\sqrt{D_0 L_\phi(w^\star)} + D_0}{2} + D_0 = \frac{\sqrt{D_0 L_\phi(w^\star)}}{2} + 2D_0 \ .$$

E Proof of Proposition 10

Proof. Let e_j's denote standard basis vectors. We have

$$\nabla_s \phi_{\mathrm{LN}}(s,y) = -\sum_{j=1}^m P_j(y) e_j + \sum_{j=1}^m \frac{\exp(s_j)}{\sum_{j'=1}^m \exp(s_{j'})} e_j$$

Therefore,

$$\begin{aligned}\|\nabla_s \phi_{\mathrm{LN}}(s,y)\|_1 &\le \sum_{j=1}^m P_j(y)\|e_j\|_1 + \sum_{j=1}^m \frac{\exp(s_j)}{\sum_{j'=1}^m \exp(s_{j'})}\|e_j\|_1 \\ &= 2.\end{aligned}$$

We also have

$$[\nabla_s^2 \phi_{\mathrm{LN}}(s,y)]_{j,k} = \begin{cases} -\frac{\exp(2s_j)}{(\sum_{j'=1}^m \exp(s_{j'}))^2} + \frac{\exp(s_j)}{\sum_{j'=1}^m \exp(s_{j'})} & \text{if } j = k \\ -\frac{\exp(s_j+s_k)}{(\sum_{j'=1}^m \exp(s_{j'}))^2} & \text{if } j \ne k \ . \end{cases}$$

Moreover,

$$\begin{aligned}\|\nabla_s^2 \phi_{\mathrm{LN}}(s,y)\|_{\infty\to 1} &\le \sum_{j=1}^m \sum_{k=1}^m |[\nabla_s^2 \phi_{\mathrm{LN}}(s,y)]_{j,k}| \\ &\le \sum_{j=1}^m \sum_{k=1}^m \frac{\exp(s_j+s_k)}{(\sum_{j'=1}^m \exp(s_{j'}))^2} + \sum_{j=1}^m \frac{\exp(s_j)}{\sum_{j'=1}^m \exp(s_{j'})} \\ &= \frac{(\sum_{j=1}^m \exp(s_j))^2}{(\sum_{j'=1}^m \exp(s_{j'}))^2} + \frac{\sum_{j=1}^m \exp(s_j)}{\sum_{j'=1}^m \exp(s_{j'})} \\ &= 2\end{aligned}$$

□